Supplementary material for Variational Inference of Population Structure in Large SNP Datasets

Anil Raj [1], Matthew Stephens [2], Jonathan K. Pritchard [3]

[1]Department of Genetics, Stanford University, Stanford, CA, 94305
[2]Departments of Statistics and Human Genetics, University of Chicago, Chicago, IL, 60637
[3]Departments of Genetics and Biology, Howard Hughes Medical Institute, Stanford, CA, 94305

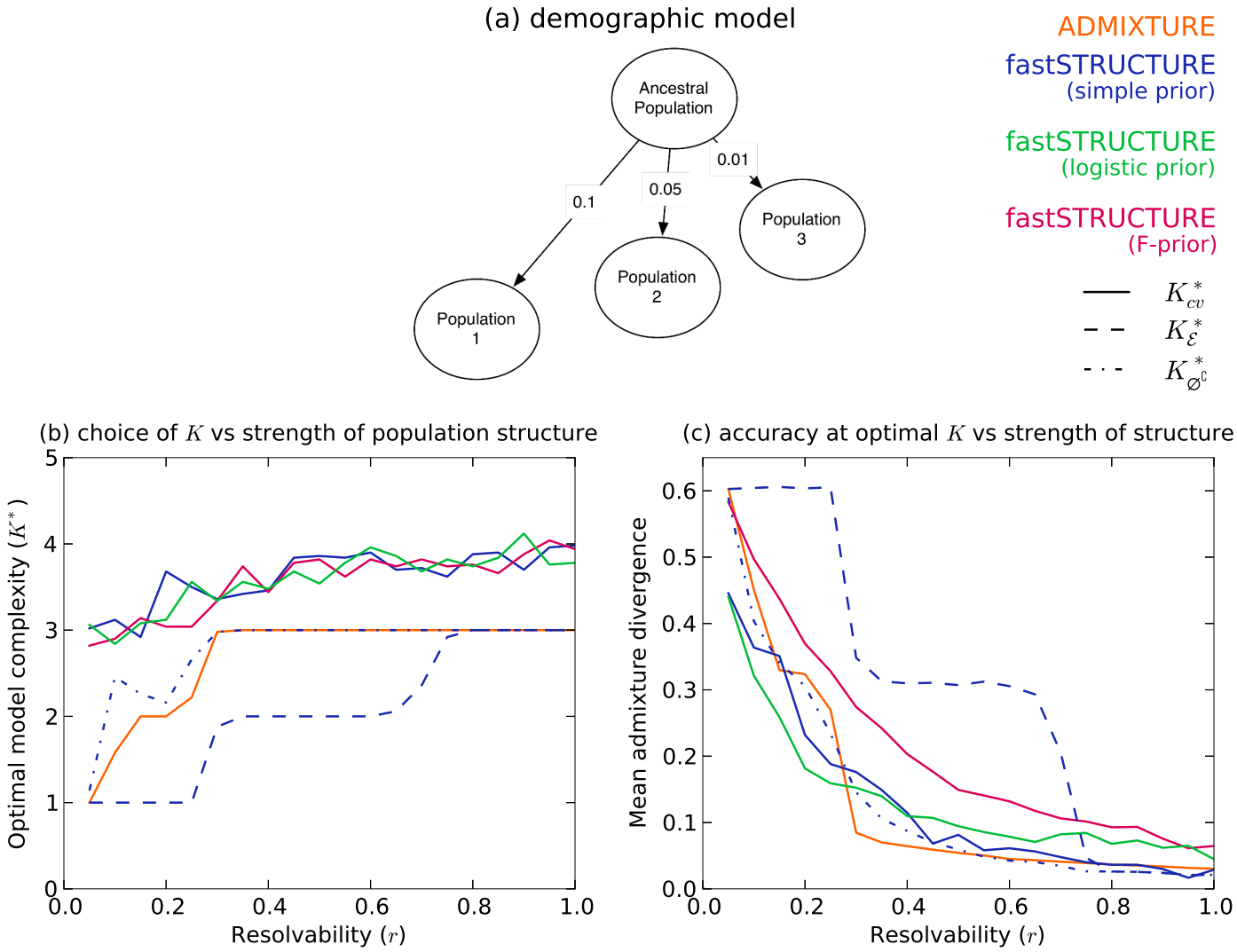

Figure S1: Accuracy of different algorithms as a function of resolvability of population structure. This figure is similar to Figure 1 in the main text, with results using the F-prior included. Subfigure (a) illustrates the demographic model underlying the three populations represented in the simulated datasets. Subfigure (b) compares the optimal model complexity inferred by ADMIXTURE (K^*_{cv}), fastSTRUCTURE with simple prior ($K^*_{cv}, K^*_{\mathcal{E}}, K^*_{\varnothing^{\complement}}$), fastSTRUCTURE with F-prior (K^*_{cv}), and fastSTRUCTURE with logistic prior (K^*_{cv}). Subfigure (c) compares the accuracy of admixture proportions estimated by each algorithm at the optimal value of K in each replicate.

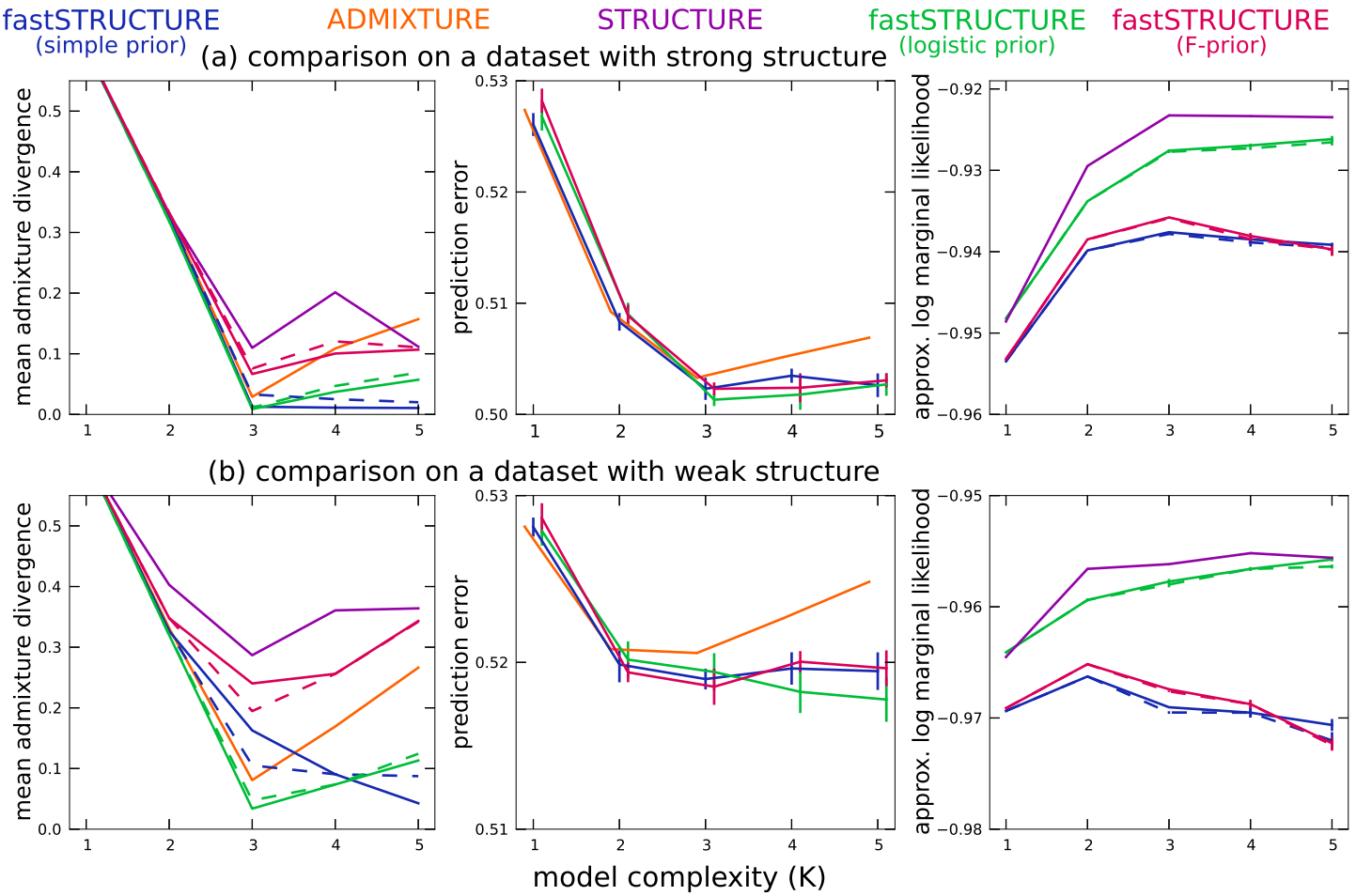

Figure S2: Accuracy of different algorithms as a function of model complexity (K) on two simulated data sets, one in which ancestry is easy to resolve (top panel; $r = 1$) and one in which ancestry is difficult to resolve (bottom panel; $r = 0.5$). This figure is similar to Figure 3 in the main text, with results using the F-prior included. Solid lines correspond to parameter estimates computed with a convergence criterion of $|\Delta\mathcal{E}| < 10^{-8}$, while the dashed lines correspond to a weaker criterion of $|\Delta\mathcal{E}| < 10^{-6}$. The left panel of subfigures shows the mean admixture divergence, the middle panel shows the mean binomial deviance of held-out genotype entries, and the right panel shows the approximations to the marginal likelihood of the data computed by STRUCTURE and fastSTRUCTURE.

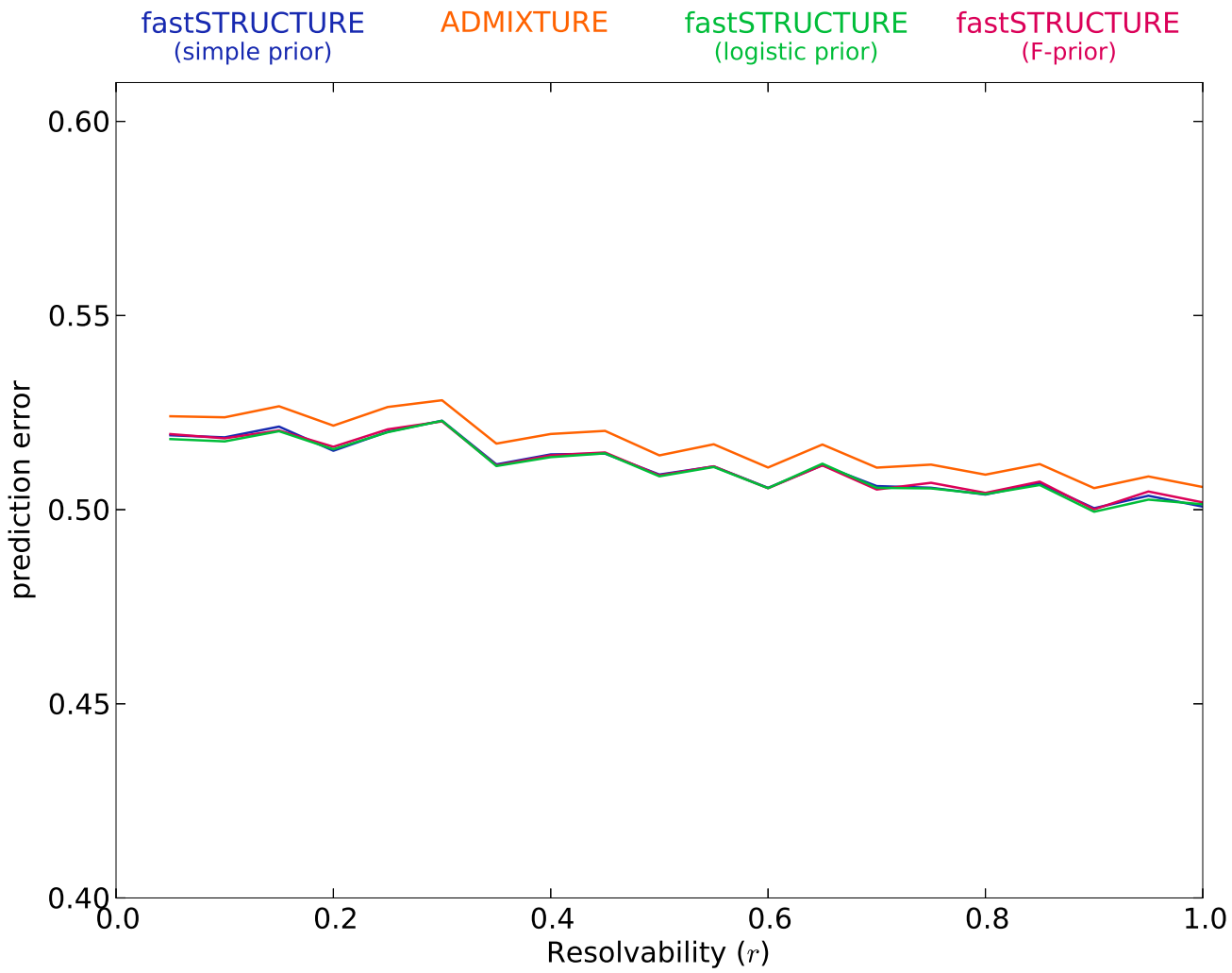

Figure S3: Prediction error of different algorithms as a function of resolvability of population structure.